ニュース記事埋め込みによる意味空間の比較研究

E5 と Qwen によるクラスタ分布および意味表現特性の比較

内山 和憲
立教大学大学院 人工知能科学研究科
24vr051p@rikkyo.ac.jp

概要

本研究では，2012 年から 2016 年の HuffPost ニュース記事 162,376 件を対象に，文埋め込みモデルが構築する意味空間の構造的差異を比較分析した。SentenceTransformer[1] を用い，E5[2] および Qwen [3] による文埋め込みを生成し、k-means クラスタリング（$k = 50$）を行い，クラスタ内分散，クラスタ重心に基づく階層クラスタリング，および社会トピックに対する時系列的反応数に着目して分析を行った。その結果，E5 は高凝集クラスタを形成し，政策・制度などの構造的トピックを明確に分離する傾向を示した。一方，Qwen は文脈的要素を反映した分散的な意味空間を形成し，社会的出来事に対する反応が複数クラスタに広く分布した。これらの結果は，文埋め込みモデルの選択が，ニュース分析における解釈の性質に本質的な影響を与えることを示唆する。

1 はじめに

近年，大規模言語モデルの発展に伴い，文章を高次元ベクトルとして表現する文埋め込みは，テキスト分析，情報検索，RAG（Retrieval-Augmented Generation）など多様な応用において重要な役割を果たしている。特に，Sentence-BERT [1] に代表される文単位の意味表現手法の登場以降，文埋め込みはトピック分析や意味的類似度計算の基盤技術として広く利用されている。

近年では，意味的近接性の学習を重視した E5 [2] や，大規模言語モデルに基づき文脈的情報を広く保持する Qwen 系埋め込みモデル [3] など，設計思想の異なる文埋め込みモデルが提案されている。

そのような文埋め込みモデルの評価を目的とした大規模ベンチマークとして，MTEB（Massive Text Embedding Benchmark）[4] が提案されており，検索や分類など多様な下流タスクにおける性能比較が行われている。

しかし，これらの評価は主として下流タスクにおける性能指標に基づくものであり，モデルごとに文書集合全体に対して構築される意味空間の構造的特性や，社会トピックに対する反応様式の差異を直接的に分析する枠組みとはなっていない。

特に、ニュース記事は社会的出来事や価値観の変化を反映するため，埋め込み空間の構造的特性は，社会トピック分析や時系列分析の解釈に直接的な影響を与える。本研究では，2012〜2016 年の HuffPost ニュース記事を対象に，E5 および Qwen の二つの文埋め込みモデルが形成する意味空間を比較し，その構造的・動的特性を明らかにすることを目的とする。

本論文の貢献は以下の三点である。

- 大規模ニュース集合に対する文埋め込み空間の凝集構造を定量的に比較した。
- クラスタ重心の階層構造に基づき，意味空間の大域的構造差を明らかにした。
- 社会トピックに対する時系列的反応様式の違いを分析した。

2 分析手法

2.1 データセット

本研究では，HuffPost が公開している HuffPost News Category Dataset[5] を用いた。同データセットには，見出し（headline），概要文（short_description），カテゴリラベル，公開日などの情報が含まれている。

時系列的な社会トピックの変化を分析するため，2012 年から 2016 年までに公開された記事のみを対象とし，合計 162,376 件の記事を抽出した。各記事について，headline と short_description を連結したテ

キストを分析対象とした。これは，記事の主題と文脈情報の双方を反映した表現を得るためである。

2.2 文埋め込みとクラスタリング

文埋め込みの生成には SentenceTransformer ライブラリを用い，E5 および Qwen の二種類の埋め込みモデルを使用した。各記事テキストを入力として，文書単位の固定長ベクトル表現を生成した。なお，生成された文埋め込みは L2 正規化されていることを確認した。

本研究で扱う短文ニュースに対して，モデルの最大トークン長による入力の切り捨てが分析結果に影響を与えていないことを確認するため，各モデルのトークナイザを用いて事前検証を行った。その結果，いずれのモデルにおいても，最大トークン長を超過する文書は存在しなかった。

得られた文埋め込みは，分析対象期間全体を通じて単一の意味空間として扱うため，全期間分を統合した上でクラスタリングを行った。クラスタリング手法としては k-means を採用し，クラスタ数は $k = 50$ と設定した。この値は，多様な社会トピックを過度に細分化せず，意味的なまとまりを保持できるバランスを考慮して決定した。

2.3 評価指標

本研究では，クラスタ内部の意味的一貫性とクラスタ間の大域的構造の双方を評価した。

まず，クラスタ内分散（Within-Cluster Variance）を用いて，各クラスタ内に含まれる文書ベクトルの分布の広がりを定量化した。以下の式で定義される。

$$\mathrm{WCV}(C_k) = \frac{1}{|C_k|} \sum_{i \in C_k} \left\| \mathbf{v}_i - \boldsymbol{\mu}_k \right\|^2 \tag{1}$$

ここで，v_i はクラスタ C_k に属する文書 i の埋め込みベクトル，μ_k はクラスタ C_k の重心ベクトルを表す。

次に，クラスタ重心間のコサイン距離に基づく階層クラスタリングを行い，意味空間の大域的構造を可視化した。さらに，特定の社会トピックに対応するクラスタ数の時系列推移を算出し，モデル間の反応様式の違いを分析した。

3 埋め込み空間の比較分析

本章では，E5 および Qwen によって形成される意味空間の違いを，クラスタ凝集性，大域的構造，および社会トピックへの反応様式の観点から分析する。

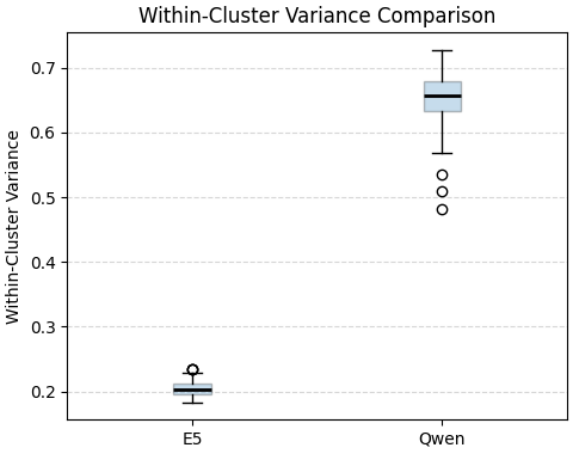

図 1 E5 および Qwen におけるクラスタ内分散の比較

3.1 クラスタ内分散の比較

図 1 は，各モデルにおけるクラスタ内分散の分布を示したものである。E5 では分散の小さいクラスタが多く，文書ベクトルが意味的に近接した高凝集クラスタを形成していることが分かる。一方，Qwen では分散が大きいクラスタが多く，クラスタ内部における意味的・文脈的多様性が高い傾向が観察された。

3.2 階層クラスタリングによる構造比較

図 2 および図 3 は，k-means により得られた $k = 50$ 個のクラスタ重心ベクトル間の距離に基づいて，階層クラスタリングを行った結果（デンドログラム）を示している。クラスタ間距離にはクラスタ重心間のコサイン距離を用い，連結法には群平均法（average linkage）を採用した。

本分析では，2012〜2016 年の米国社会において継続的に議論されていた主要な社会的関心領域として、「政治・選挙（Politics & Elections）」，「LGBT・市民権（LGBT & Civil Rights）」，「公民権・人種問題（Civil Rights & Race）」，「気候・環境（Climate & Environment）」，「政治・公共安全（Politics & Public Safety）」の 5 つの社会トピックを対象とした。

各クラスタには，これら 5 つの社会トピックに基づく色分けを施している。クラスタのトピック識別は，あらかじめルールベースにより社会トピックラベルが付与された記事集合を用い，当該クラスタに含まれる記事の中で出現頻度が最も高いトピックを

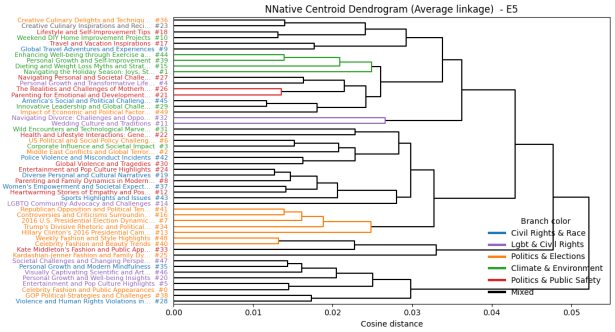

図 2　階層クラスタリング結果（E5）

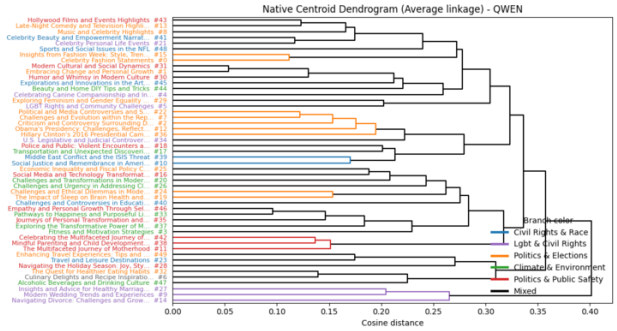

図 3　階層クラスタリング結果（Qwen）

そのクラスタの代表トピックとして採用する多数決方式によって行った。

具体的には，記事レベルでは各社会トピックに対して，あらかじめ定義したキーワード集合に基づくルールベースの判定を行った。例えば，政治・選挙トピックでは election，vote，republican，democrat，trump，obama などの語を含むキーワード集合を用いている。各トピックに対応するイベントおよびキーワード集合の詳細は付録に示す。

記事本文（headline と short_description の連結文）中に，これらのキーワードが一定数以上（本研究では 2 語以上）出現した場合に，当該社会トピックのラベルを付与した。このルールに基づき，社会トピックに該当すると判定された記事数は $N_{\text{label}} = 8{,}753$ 件である。これらの記事を用いて，各クラスタ内におけるトピック別記事数を集計し，最も多くの記事を含むトピックをそのクラスタの代表トピックとして決定した。その結果，いずれのモデルにおいても，全クラスタがいずれかの社会トピックに一意に割り当てられることを確認した。

E5 のデンドログラムでは，政治・選挙，公民権・人種問題，気候・環境といった主要な社会トピックが比較的明確な枝として分離されており，同一トピックに属するクラスタが階層構造上でも近接して配置される傾向が確認された。これは，E5 によって構築される意味空間が，主題的観点に基づいて整理された大域的構造を持つことを示唆している。

一方，Qwen のデンドログラムでは，異なる社会トピックに属するクラスタが同一の枝に混在する場合が多く観察された。この結果は，Qwen の意味空間が，明示的な主題境界よりも文脈的・叙述的近接性を重視して構成されており，複数の社会トピックを横断的に結び付ける連続的な構造を持つことを示している。

3.3　社会トピック反応数の時系列分析

本節では，第 3.2 節で扱った社会トピックに対して、E5 および Qwen によって形成されたクラスタが，時系列的にどの程度反応しているかを比較分析する。

本節における「反応」とは，第 3.2 節におけるクラスタ代表トピックの一意決定とは異なり，社会トピックに関連する記事がクラスタ空間内にどのように分布しているかに着目した概念である。本分析は，クラスタ全体の主題的性質を定めることを目的とせず，社会トピックへの関心の広がりを時系列的に把握することを目的とする。

具体的には，各記事本文（headline と short_description の連結文）に対して，第 3.2 節と同一のトピック別キーワード集合を用いたルールベースの判定を行い，同一トピックに属するキーワードが一定数以上（本研究では 2 語以上）出現した場合に，当該記事をその社会トピックに反応したものと判定した。その上で，各四半期において，当該トピックに反応した記事を 1 件以上含むクラスタ数を「トピック反応クラスタ数」と定義し，モデル別に集計した。

本分析では，クラスタを排他的な主題単位として扱うのではなく，社会トピックに対する関心の広がりや集中の度合いを補助的に捉える。

図 4〜図 6 は，上記 5 つの社会トピックについて，各四半期におけるトピック反応クラスタ数の差分（E5 − Qwen）を示している。

全体として，政治・選挙および政治・公共安全といった出来事性の高いトピックでは，特定の時期に Qwen 側で反応クラスタ数が増加する傾向が観察された。同一の社会トピックであっても，異なる文脈や論点に基づく記事が，単一のクラスタに集約されにくい傾向があると解釈できる。一方，E5 では，同

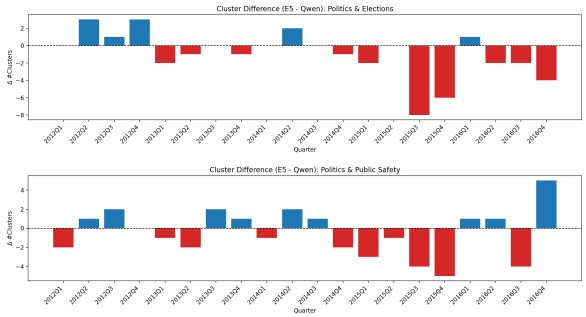

図 4　政治・選挙および政治・公共安全トピックのクラスタ反応数差分

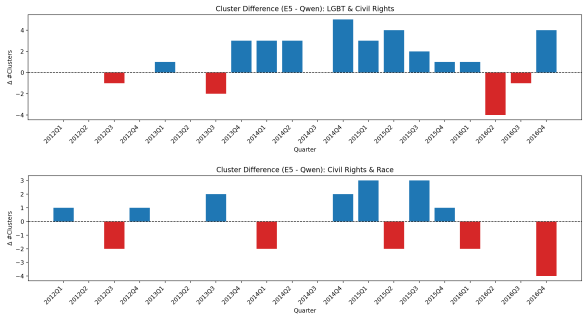

図 5　LGBT・市民権および公民権・人種問題トピックのクラスタ反応数差分

様のトピックにおいても，反応が比較的限定されたクラスタに集中しており，出来事を制度的・政策的枠組みの中で整理する方向の表現が反映されていると考えられる。

また，LGBT・市民権，公民権・人種問題，および気候・環境といったトピックでは，E5 側で一貫して安定した反応クラスタ数が観察された。これらのトピックは，長期的な制度議論や社会構造と結びついた話題を含むことが多く，E5 のクラスタ空間内では，関連する記事が比較的明確な主題的まとまりを保った形で分布していることが示唆される。一方，Qwen では，同一トピックに関連する記事がより広い範囲のクラスタに分散して反応しており，同一の社会トピックが複数の視点や文脈を通じて表現されている可能性が示されている。

小括　以上の時系列分析から，E5 は制度，政策，および長期的社会課題に関連するトピックに対して，比較的安定したクラスタ反応を示す一方で，Qwen は特定の時期において，関連する記事がより多くのクラスタに分散して反応する傾向を持つことが確認された。この結果は，第 3.2 節で示した意味空間構造の違いを，時系列的な反応分布という観点から補完するものである。

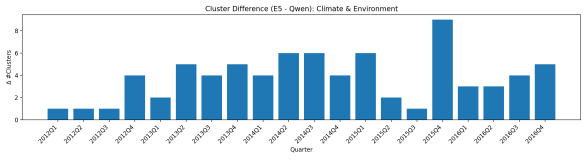

図 6　気候・環境トピックのクラスタ反応数差分

4　考察

本研究の結果は，文埋め込みモデルの設計思想の違いが，ニュース記事集合に対する意味空間構造として一貫して反映されることを示している。

E5 はクラスタ内分散が小さく，意味的に近接した記事を安定して集約する高凝集なクラスタを形成した。また，階層クラスタリングにおいても，政治や公民権，気候・環境といった社会トピックが比較的明確に分離される傾向が確認された。この特性は，主題分類や意味検索など，意味境界の明確さが求められる応用において有効であると考えられる。

一方，Qwen はクラスタ内分散が大きく，文脈的近接性を反映した分散的なクラスタ構造を示した。社会トピック反応数の時系列分析からは，特定の出来事に関連する記事が複数のクラスタに分散して反応する傾向を示した。これは，社会的関心の広がりや変動が，意味空間上で分散的に表現される特性を示唆している。

以上より，文埋め込みモデルの選択は，単なる性能指標の比較に留まらず，分析対象とする社会現象の性質や求められる解釈の粒度に応じて行う必要があることが示された。

特に，ニュース記事のように社会的出来事と長期的課題が混在するテキスト集合においては，埋め込みモデルが形成する意味空間の構造そのものを理解することが重要である。本研究の結果は，下流タスクの性能評価のみでは捉えにくい，文埋め込みモデル固有の意味表現特性を可視化する一つの分析視点を提供する。

5　おわりに

本研究では，ニュース記事を対象に文埋め込みモデル間の意味空間構造の違いを比較した。今後は，他モデルや外部社会データとの統合を通じて，さらなる分析を行う予定である。

参考文献

[1] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (arXiv:1908.10084), 2019.

[2] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text Embeddings by Weakly-Supervised Contrastive Pre-training (arXiv:2212.03533), 2024.

[3] Jinze Bai, et al. Qwen Technical Report (arXiv:2309.16609), 2023.

[4] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive Text Embedding Benchmark (arXiv:2210.07316), 2023.

[5] Rishabh Misra. News Category Dataset (arXiv:2209.11429), 2022.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

A　社会的トピックに対応するイベントおよびキーワード一覧

本付録では，本文において用いた社会的トピック判定に対応する，代表的なイベントおよびキーワード集合の全体像を示す。記事レベルのトピック判定は，これらのキーワード集合に基づいて行われている。なお，一部の代表的事件については，表記揺れによる検出漏れを防ぐため，辞書に含まれるキーワード集合を保持したまま，同義的表現を補助的に追加した。

表 1: 社会的トピックに対応するイベントおよびキーワード一覧

Topic Group	**Event Name**	**Keywords**
Civil Rights & Race	Trayvon Martin shooting	Trayvon Martin; George Zimmerman; Stand Your Ground; racial profiling; gun violence
Civil Rights & Race	Black Lives Matter founding	Black Lives Matter; BLM; police brutality; racial justice; protest movement
Civil Rights & Race	Eric Garner death (“I can’t breathe”)	Eric Garner; I can’t breathe; NYPD; police chokehold; Staten Island
Civil Rights & Race	Ferguson unrest	Ferguson; Michael Brown; Darren Wilson; Ferguson protests; police shooting
Civil Rights & Race	Baltimore unrest (Freddie Gray)	Freddie Gray; Baltimore riots; police custody death; protest; unrest
Politics & Public Safety	Sandy Hook shooting	Sandy Hook; Newtown; school shooting; gun control; mass shooting
Politics & Public Safety	Orlando Pulse nightclub shooting	Orlando Pulse; nightclub shooting; mass shooting; gun violence; terrorism
Politics & Elections	2012 U.S. Presidential Election	Obama Romney; 2012 election; presidential debate; swing states; voters
Politics & Elections	2016 U.S. Presidential Election	Donald Trump; Hillary Clinton; 2016 election; campaign; populism
Climate & Environment	Paris Climate Agreement	Paris Agreement; climate change; COP21; global warming; emissions
Climate & Environment	Deepwater Horizon aftermath	Deepwater Horizon; BP oil spill; Gulf of Mexico; environmental damage
LGBT & Civil Rights	U.S. Supreme Court same-sex marriage ruling	same-sex marriage; Obergefell v. Hodges; Supreme Court; LGBT rights